WWW GUIDE

Nucleic acid databases on the Web

Richard Peters and Robert S. Sikorski

For many researchers, searchable sequence databases on the Internet are becoming as indispensable as pipettes. Especially in the fields of molecular and structural biology, many excellent online databases have cropped up in just the past year. Despite this wealth of information, it still requires considerable effort to sort through these myriad databases to find ones that are useful. Keeping this in mind, we started by using our Net search tools to see if we could winnow out the best of the Net. This month, we decided to focus on nucleic acid databases (protein databases will appear next month).

The richness of the sites we have listed is really extraordinary. An outstanding example is The National Center for Biotechnology Information (NCBI) site. Some of the things you can now find at their site include BLAST, Genbank, Entrez, dbEST, dbSTS, OMIM, and UniGene. A brief description of each of these tools demonstrates what we mean. BLAST is the standard program used for querying your DNA or protein sequence against all known sequences housed in databases. You can even have the output sent to you by e-mail. Genbank is an annotated collection of pretty much all publicly available DNA sequences. It forms the "known" database for BLAST searches. Entrez is a window into the molecular biology subset—about 1/7—of MEDLINE, combined with access to nucleic acid and protein databases (GenBank, EMBL, DDBJ, PIR, SWISS-PROT, PRF, and PDB—more on these databases next month). dbEST is a database of expressed sequence tags (ESTs) that represent short segments of transcribed genes. These ESTs come from both the general public and the Washington University (St. Louis, MO) sequencing project. dbSTS is a database of genomic sequence tagged sites that serve as mapping landmarks in the human genome. OMIM (Online Mendelian Inheritance in Man) is a unique database that catalogs current research and clinical information about individual human genes. UniGene is a subset of GenBank that defines a collection of DNA sequences likely to represent the transcription products of distinct genes. The UniGene entries are being physically mapped to the human genome to create a "transcript map."

Recently, GenBank, EMBL, and DDBJ have developed shared rules to describe the roles and locations of higher order sequence domains and elements within the genome of an organism. This architecture allows for simplified exchange of data between these three large databases so that, conceivably, one should be able to access all the data from these three databases without having to access each of them separately.

Richard Peters is at the department of medicine, Masachusetts General Hospital and Robert S. Sikorski is a Howard Hughes Medical Institute fellow at the National Cancer Institute (rpeters@vp3.med.harvard.edu; http://www.-medsitenavigator.com and rss@nchgr.nih.gov).

A sampler of products and services sites related to nucleic acid databases

Codon usage database	http://www.dna.affrc.go.jp/nakamura/	An amazing database of the codon usage of some 5163 organisms.
Databases at the EBI	http://www.ebi.ac.uk/dbases/topdata.html	Databases at the European Bioinformatics Institute for nucleotide protein searches. (Data largely overlap that at NCBI.)
DNA data bank of Japan	http://www.ddbj.nig.ac.jp/	An interface to the a nucleic acid sequence db maintained in Japan.
Human genome centers	http://www-hgc.lbl.gov/inf/HGcenters.html	A listing of the world's human genome centers. Each center contains their own unique collections of databases.
Mitomap v. 3.0	http://www.gen.emory.edu/mitomap.html	A clearinghouse for mitochondrial genomics that tries to link mitochondrial gene structure to function.
Mouse genome database	http://www.informatics.jax.org/mgd.html	A major collection of mouse genetic and phenotypic data.
NCBI	http://www.ncbi.nlm.nih.gov/	Home of BLAST, GenBank, Entrez, dbEST, dbSTS, UniGene.
OMIA database query form online	http://www.angis.su.oz.au/BIRX/omia/omia_form.html	Mendelian Inheritance in Animals is an interesting collection of animal diseases, some with similarities to those found in humans.
OMIM	http://www3.ncbi.nlm.nih.gov/omim/	Catalog of human genes and genetic disorders
Saccharomyces genome database	http://genome-www.stanford.edu/Saccharomyces/	The culmination of the yeast genome project.
The genome database	http://gdbwww.gdb.org	A focal database for human gene mapping that attempts to integrate physical and genetic maps.
TIGR human cDNA database	http://www.tigr.org/tdb/hcd/overview.html	Billed as the world's largest cDNA database, this db is accessible online, but its use comes with some restrictions.
Vector sequence database	http://biology.queensu.ca/miseners/vector.html	An impressive collection of sequence and other information about the most commonly used cloning vectors.